PREDICTING DIABETES USING THE PIMA INDIAN DIABETES DATASET OF AKIMEL O' ODHAM HERITAGE

alex adenuga

2023-01-28

Background

This report explored the Pima Indian Diabetes data set of the people of Gila River Indian Community where all patients are females who are at least 21 years old. It was extracted as a sub-data that was collected from the National Institute of Diabetes and Digestive Kidney Diseases (NIDDK) data derived from an epidemiological research done between 1960s and 1970s. NIDDK is a subsection of National Institute of Health in the United States. The methodologies used are logistic regression model and Naive Bayes Models; these are classification models for binary response variables.

The analytical processes followed when analyzing the data showed the guidelines required to predict futuristic incidences of having diabetes when using classification models. A comparison between the Logistic regression model and Naive Bayes model was performed using 0.5 and 0.25 threshold for prediction of probability that a patient has diabetes or not. We examined the the error rate and the changes that occurred in the confusion matrix in each and both models.

***Methods and Code chunks**

The logistic regression model and the Naive Bayes modelling methods were used in this report.

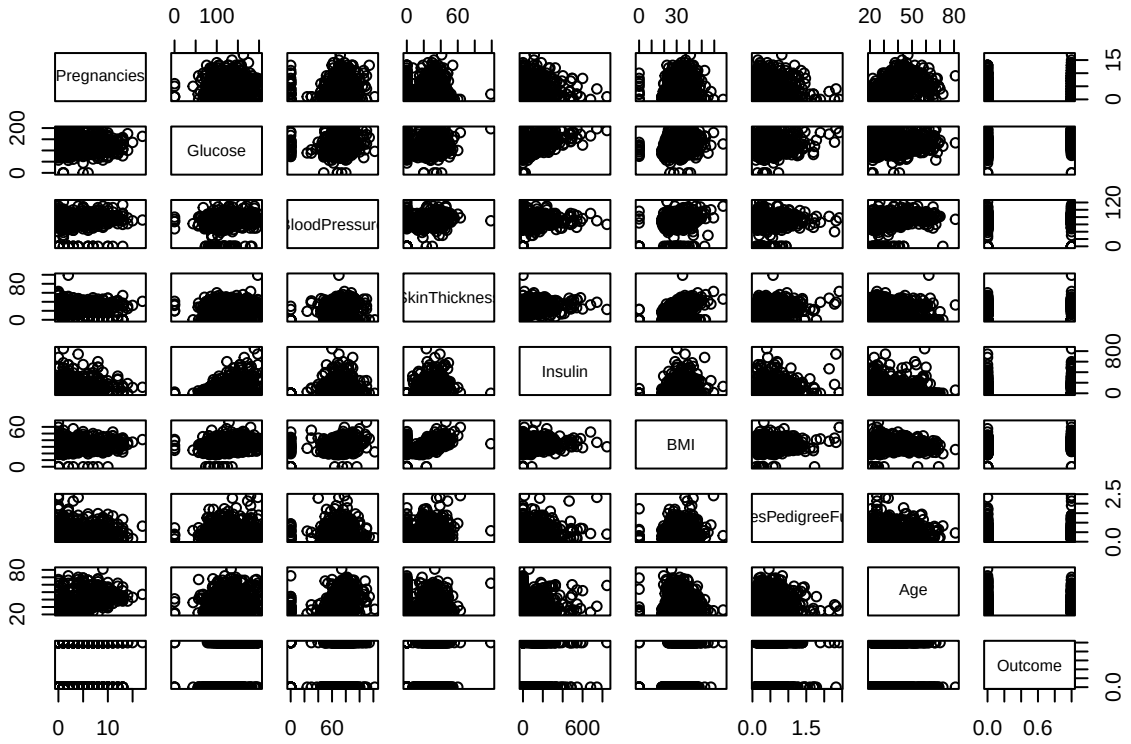
##The data contained:

```
diabetes<-read.csv("diabetes.csv")
names(diabetes)
```

```
## [1] "Pregnancies"              "Glucose"
## [3] "BloodPressure"            "SkinThickness"
## [5] "Insulin"                  "BMI"
## [7] "DiabetesPedigreeFunction" "Age"
## [9] "Outcome"
```

#Visualizing the data

```
plot(diabetes)
```



Note: There seem to be correlation between Pregnancies and Age. correlation between Insulin and SkinThickness

#checking the correlation to be sure of what showed on the plot

```
cor(diabetes)
```

Note: The correlation matrix showed that Pregnancies and Age has correlation of (0.54). While Insulin and SkinThickness has a correlation of (0.44). The implication is that we may decide to remove one or both pairs of variables from the model.

#The Logistic Regression Model

```
mod_log<-glm(Outcome~.-Age-SkinThickness-Insulin, data = diabetes, family = binomial)
summary(mod_log)
```

```
##
## Call:
## glm(formula = Outcome ~ . - Age - SkinThickness - Insulin, family = binomial,
##     data = diabetes)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -7.954952   0.675823 -11.771  < 2e-16 ***
## Pregnancies               0.153492   0.027835   5.514  3.5e-08 ***
## Glucose                   0.034658   0.003394  10.213  < 2e-16 ***
## BloodPressure            -0.012007   0.005031  -2.387  0.01700 *
```

```
## BMI                       0.084832   0.014125   6.006  1.9e-09 ***
## DiabetesPedigreeFunction  0.910628   0.294027   3.097  0.00195 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 728.56  on 762  degrees of freedom
## AIC: 740.56
##
## Number of Fisher Scoring iterations: 5
```

Note: Due to the non significant behavior of some of the variables; possibly as a result of confounding; Age, SkinThickness and Insulin were removed from the model to improve the model performance.

#The Logistic regression model Prediction

```
Pred_diabetes<-predict.glm(mod_log, type = "response")
Pred_diabetes[1:10]
```

```
##          1          2          3          4          5          6          7
## 0.65750317 0.04428403 0.80775101 0.04863693 0.88621766 0.15434062 0.07338081
##          8          9         10
## 0.66423427 0.74467399 0.03435068
```

#Logistic regression Model Accuracy Test with 0.5 threshold

```
Test_pred_diabetes<-ifelse(test = Pred_diabetes>0.5,
                           yes = 1,
                           no = 0)
Tab_pred_diabetes<-table(diabetes$Outcome, Test_pred_diabetes)

Tab_pred_diabetes
```

```
##    Test_pred_diabetes
##       0   1
##   0 441  59
##   1 114 154
```

```
ErrorRate<-(Tab_pred_diabetes[1,2] + Tab_pred_diabetes[2,1]) / sum(Tab_pred_diabetes) * 100

ErrorRate
```

```
## [1] 22.52604
```

#Logistic regression Model Accuracy Test with 0.25 threshold

```
Test_pred_diabetes2<-ifelse(test = Pred_diabetes>0.25,
                           yes = 1,
                           no = 0)
Tab_pred_diabetes2<-table(diabetes$Outcome, Test_pred_diabetes2)

Tab_pred_diabetes2
```

```
##    Test_pred_diabetes2
##       0   1
##   0 317 183
##   1  46 222
```

```
ErrorRate2<-(Tab_pred_diabetes2[1,2] + Tab_pred_diabetes2[2,1]) / sum(Tab_pred_diabetes2) * 100

ErrorRate2
```

```
## [1] 29.81771
```

#Naive Bayes Model

```
library(e1071)
set.seed(100)

mod_nb<-naiveBayes(Outcome~.-Age-SkinThickness-Insulin, data = diabetes)

Pred_diabetes_nb<-predict(mod_nb, newdata = diabetes, type = "raw")[, 2]

Pred_diabetes_nb[1:10]
```

```
##  [1] 0.57494714 0.03180799 0.86674398 0.03818108 0.99974901 0.09231473
##  [7] 0.04861702 0.83196538 0.89085486 0.01687940
```

#Naive Bayes Model Accuracy Test with 0.5 threshold

```
Test_pred_diabetes_nb<-ifelse(test = Pred_diabetes_nb > 0.5,
                              yes = 1,
                              no = 0)
Tab_pred_diabetes_nb<-table(diabetes$Outcome, Test_pred_diabetes_nb)

Tab_pred_diabetes_nb
```

```
##    Test_pred_diabetes_nb
##       0   1
##   0 438  62
##   1 110 158
```

```
ErrorRate_nb<- (Tab_pred_diabetes_nb[1,2] + Tab_pred_diabetes_nb[2,1]) / sum(Tab_pred_diabetes_nb) * 10

ErrorRate_nb
```

```
## [1] 22.39583
```

#Naive Bayes Model Accuracy Test with 0.25 threshold

```
Test_pred_diabetes_nb2<-ifelse(test = Pred_diabetes_nb > 0.25,
                              yes = 1,
                              no = 0)
Tab_pred_diabetes_nb2<-table(diabetes$Outcome, Test_pred_diabetes_nb2)

Tab_pred_diabetes_nb2
```

```
##    Test_pred_diabetes_nb2
##       0   1
##   0 357 143
##   1  58 210
```

```
ErrorRate_nb2<- (Tab_pred_diabetes_nb2[1,2] + Tab_pred_diabetes_nb2[2,1]) / sum(Tab_pred_diabetes_nb2) *

ErrorRate_nb2
```

```
## [1] 26.17188
```

Discussion

Looking at the logistic regression model, It was observed that a reduction in the prediction threshold from to lead to in increase in the false positive prediction from 59 to 183, a reduction in the false negative from 114 to 46, a decrease in the true negative from 441 to 317 and an increase in the true positive from 154 to 222; noting that there was also an increase the the error rate.

In the same pattern, the Naive Bayes model recorded there was an increase in the false positive prediction from 62 to 143, a reduction in the false negative from 110 to 58, a decrease in the true negative from 438 to 357 and an increase in the true positive prediction from 158 to 210; noting that there was also an increase in the error rate from 22.4% to 26.1%

The 0.5 and 0.25 threshold was used to test the model accuracy of the Logistic regression model and the Naive Bayes model.It was noted that the Naive Bayes Model has the smallest error rate with 22.3% upon prediction while the logistic regression model had 22.5% using greater than 0.5 as the predictions threshold for the probability of having diabetes and 26.2% and 29.8% respectively when the prediction threshold was changed to 0.25.

Conclusion

Comparing the prediction performance of Logistic regression model and Naive Bayes model using the diabetes data: The Naive Bayes model appear to perform better using the error rate as it has a lower error rate upon prediction even when the prediction threshold was changed. It is worthy of note that the behavior of both models were identical when the prediction threshold was changed from 0.5 to 0.25.